ISSN 1539–2791

Volume 1 • Number 4 • 2003

Neuroinformatics

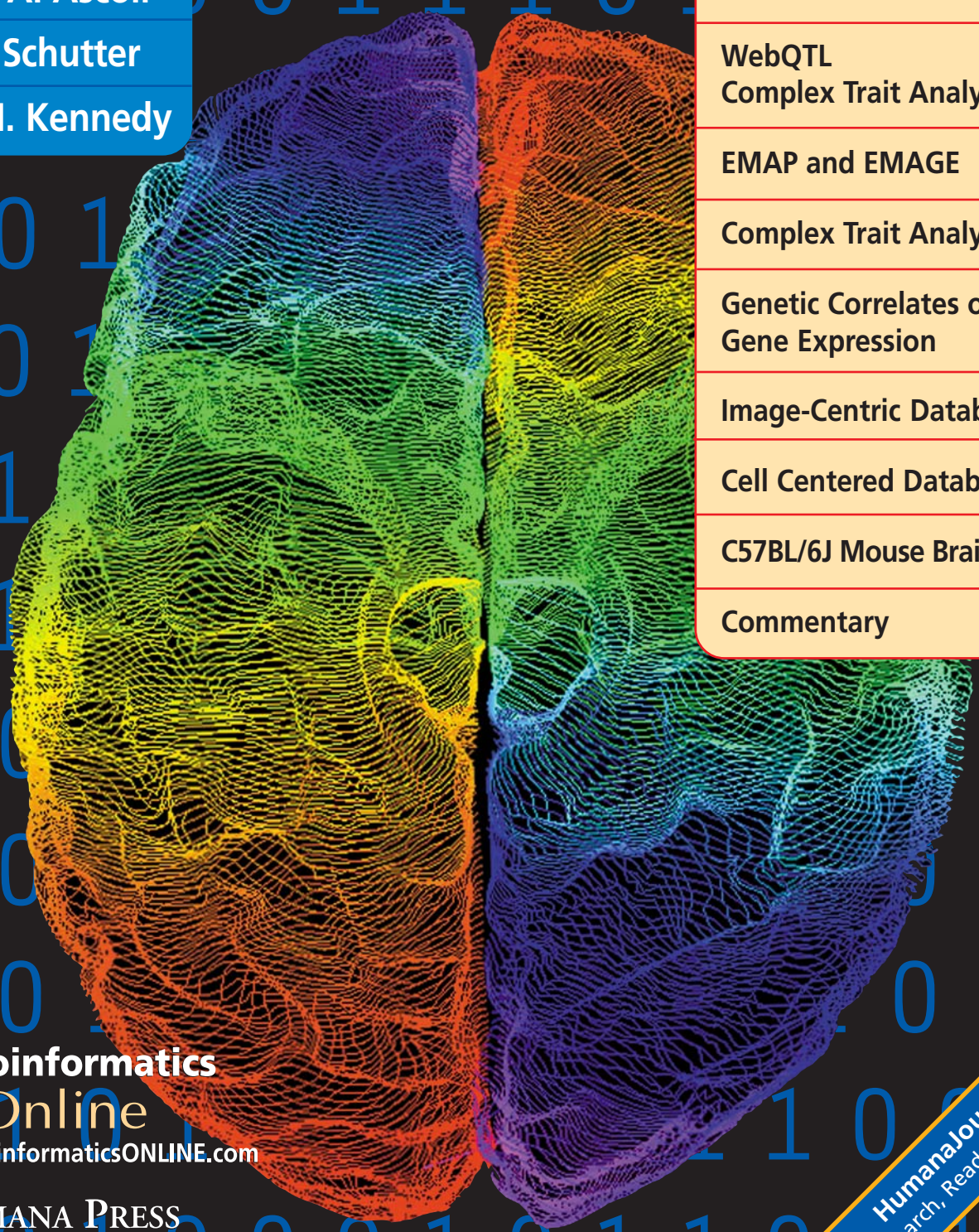
Editors

Giorgio A. Ascoli

Erik De Schutter

David N. Kennedy

IN THIS ISSUE



Neuroinformatics
Online
www.NeuroinformaticsONLINE.com

HUMANA PRESS

HumanaJournals.com
Search, Read, and Download

Neuroinformatics
Copyright ©Humana Press Inc.
All rights of any nature whatsoever are reserved.
ISSN 1539-2791/03/299–308/$25.00

Original Article

WebQTL

Web-Based Complex Trait Analysis

Jintao Wang,[1] Robert W. Williams,[2] and Kenneth F. Manly[*,1]

[1]Department of Molecular & Cellular Biology, Roswell Park Cancer Institute, Buffalo, NY; [2]Department of Anatomy and Neurobiology, Center of Genomics and Bioinformatics, University of Tennessee Health Science Center, Memphis, TN

Abstract

WebQTL is a website that combines databases of complex traits with fast software for mapping quantitative trait loci (QTLs) and for searching for correlations among traits. WebQTL also includes well-curated genotype data for five sets of mouse recombinant inbred (RI) lines. Thus, to identify QTLs, users need provide only quantitative trait data from one of the supported populations. The WebQTL databases include both biological traits—neuroanatomical, pharmacological, and behavioral traits—and microarray-based gene expression data from BXD RI lines. A search function finds correlations between RNA expression and biological traits, and mapping functions find QTLs for either type of trait. The WebQTL service is available at http://www.webqtl.org/.

Index Entries: Genetics; genetic map; quantitative trait; complex trait; microarray; transcriptome; software; web service.

Introduction

The genetic architecture of the brain and behaviors is complex and involves variation at many genes, interactions between those genes, response to environment, gene-by-environment interactions, and results of stochastic events. In order to facilitate analysis of such multigenic traits under different conditions it is essential to use a common reference panel of isogenic strains that can be systematically phenotyped and analyzed in multiple ways. WebQTL is a novel Internet resource that combines analysis software with key data for a common panel of over 100 recombinant inbred (RI) strains, all of which are commercially available from The Jackson Laboratory (Bar Harbor, ME).

RI lines are sets of inbred lines derived by sib-mating of the progeny of a genetic cross; each set is a permanent resource providing unlimited individuals representing the same set of genetic recombination events. Using these

*Address to which all correspondence and reprint requests should be sent.
E-mail: kenneth.manly@roswellpark.org

resources, complex traits can be analyzed by searching for statistical associations among complex traits or searching for associations between a trait and genotypes at known marker loci. The latter searches define quantitative trait loci (QTLs), regions expected to contain genes controlling the trait (Broman, 2001; Mackay, 2001; Barton and Keightley, 2002; Doerge, 2002; Phillips and Belknap, 2002). This combination of analysis software and data resources enables new types of complex trait analysis (Bystrykh et al., 2003; Chesler et al., 2003).

Web-Based Complex Trait Analysis

Websites are a relatively recent method for distributing computational tools for complex trait analysis. They offer certain advantages for this purpose. First, they are available to any computer with a web browser, although differences between browsers can affect function significantly. Second, users always benefit from the latest revision of the software without explicit installation of updates. Third, the website can provide information needed for analysis, such as genotypes for RI lines or map positions for commonly used genetic markers. Finally, analysis services can be combined with links to related information available on the web. These possibilities are realized to varying degrees by currently available QTL Web services.

QTL Café

QTL Café (Seaton et al., 1998) was the first QTL mapping web service to be developed. It consists of a Java applet, available at http://web.bham.ac.uk/g.g.seaton/, which can be run in a suitable web browser. QTL Café offers single-marker analysis of variance, simple interval mapping by regression, and marker-difference regression mapping. It supports backcross, intercross (F2), selfed RI, and double haploid designs.

QTL Café requires input data in three text files: a file with map information for marker loci, a file with names of traits and trait values, and a genotype file with genotypes of marker loci. Output is in the form of graphs and tabular data.

QTL Express

QTL Express (Seaton et al., 2002) differs from QTL Café in that it provides QTL analysis tools for outbred populations (which also work for inbred populations). The site, available at http://latte.cap.ed.ac.uk/, provides methods for backcross populations, F2 populations, mixtures of backcross and F2 populations, collections of half-sib families, and collections of sib-pairs. The method for F2 populations is the Haley and Knott (1992) regression method as extended for outbred populations (Haley et al., 1994). The method for half-sib families is that developed by Knott, Elsen, and Haley (1996).

QTL Express reads data from text files in three formats. The first is a genotype file that provides the names of marker loci and sex, pedigree, and genotype information for each individual, including grandparents and F1 parents, if available. The second file is a map file that provides the order and position information for markers of each chromosome (with sex-specific map distances, if available). The third file provides values for quantitative traits, fixed effects, and covariates.

The initial Web page at the QTL Express site accepts file names for the three data files. When these are submitted, QTL Express calculates the genotype probabilities at each locus for each individual and returns a page that summarizes this information. A subsequent page then allows the user to choose either a one- or two-QTL model with cofactors and/or interactions and to specify other evaluation parameters. The results page summarizes the evidence for QTLs with both summary tables and graphs of variance ratios and regression coefficients.

HAPPY

HAPPY (Mott et al., 2000; Mott and Flint, 2002), available at http://www.well.ox.ac.uk/happy/, is designed to map QTL in heterogeneous stocks, populations derived from multiple inbred lines that have interbred for many generations. The advantage of this type of population is the potential mapping precision, which is the result of many generations of accumulated recombination.

HAPPY uses allele information from the founder lines, but does not use pedigree information for the individuals in the mapping population (which, if available, would be very complex). It requires two input text files. The first defines the founder strains and the alleles they each carry at all marker loci. The second lists the trait value and marker genotypes for all individuals in the mapping population.

Analysis by HAPPY is a two-step process. In the first step, HAPPY uses a multipoint method that combines genotypes and founder haplotypes to estimate the probability of each founder being the ancestor of a given allele in a progeny individual. In the second step, the association between trait values and these allele probabilities is evaluated, using a model that assumes that the trait is determined by additive contributions from each of the two alleles at a locus. The output of HAPPY is a file of tabular data estimating for each interval 1) the probability of each founder line contributing to the genotype, 2) an estimate of the QTL contribution for each founder, and 3) an F-statistic evaluating the significance of the trait-marker association in that interval.

WebQTL: Beyond QTL Mapping

Like the software described previously, WebQTL will search for QTLs, but it is designed specifically to support shared mapping populations—five sets of mouse RI lines and one advanced intercross. That is, WebQTL will search for QTLs using traits measured in the CXB, AXB, BXA, BXD, or BXH sets of mouse RI lines (Williams et al., 2001). RI lines offer some advantages for this type of analysis. They are shared mapping resources in the sense that they are commercially available and, because they are inbred, experiments done in different laboratories are directly comparable. In addition, genotypes for RI strains at hundreds of loci are known and publicly available. The availability of unlimited numbers of genetically identical individuals from each line offers the chance to reduce trait sampling error. Although the number of lines in each mouse RI set is limited, this limitation can be overcome by the recombinant inbred intercross (RIX) strategy, a strategy that uses defined hybrids by combinatorial mating of the RI lines (Threadgill et al., 2001).

WebQTL is more than a mapping engine, however, because it includes integrated databases with both genotype and quantitative trait data for the supported RI lines, particularly the BXD set. These data resources are described more fully below. Finally, WebQTL will search for correlations between a selected or user-defined trait and traits in one of its databases. This feature is especially useful for finding correlations between biological traits and expression of particular genes as assayed by microarray hybridization (Chesler et al., 2003). Figure 1 shows an example of such a search of the database of published biological traits for those correlated with expression of the Shh (sonic hedgehog) gene.

The correlation with hindbrain weight is compatible with the recent demonstration of the effect of sonic hedgehog on neural progenitor proliferation (Lai et al., 2003).

QTL Mapping

Complex trait data can be submitted by a user or retrieved from one of the integrated databases. Figure 2 shows an example of a query form for retrieving data for gene expres-

Correlation Results

Trait (ProbeSet/101831_at) values were compared to all values in **Published Phenotypes** database. The TOP 100 correlations (Pearson linear correlation coeffcient, absolute value) are displayed below.The p-value shown is a comparison-wise error rate (CWER), uncorrected for multiple comparisons.

Clicking on the record ID will open the published phenotype data for that publication. Click on the correlation to see a scatter plot of the trait data. Click on the URL to see the pub med abstract for your trait.

☑ Display strain names in correlation plot
☑ Display fit line in correlation plot

	Record ID	Phenotype	Authors	Year	URL	Correlation	#Strains	p Value
1	**0.81**	hindbrain weight	Williams RW unpub	2000	N/A	0.7402	16	0.0006
2	**10371755.04**	acoustic startle response to 20 kHz white-noise bursts at 100 dB	McCaughran J, Bell J, Hitzemann R	1999	Pub Med	-0.6795	23	0.0002
3	**8430813.01**	lavageable bronchoalveolar polymorphonuclear leukocytes (PMNs) 6h after acute (3h) exposure to 2ppm ozone	Kleeberger SR, Levitt RC, Zhang LY.	1993	Pub Med	-0.6719	17	0.0023
4	**10371755.06**	acoustic startle response to a 110 dB SPL, 10 kHz	McCaughran J, Bell J, Hitzemann	1999	Pub Med	-0.6688	23	0.0003

Fig. 1. Results of a correlation search, showing part of the list of correlations between Shh RNA levels and behavioral and physiological traits.

sion in brain. Whether user-submitted or retrieved from the database, WebQTL displays the data in a table that can be checked for accuracy and edited. This page then remains open to control further analysis while results of different procedures appear in new pages (Fig. 3).

Often, the first step in analysis of new trait data is single-marker regression across all chromosomes. A hypothetical QTL is evaluated at the location of each marker locus, and the significance of that QTL is estimated from a likelihood ratio statistic (LRS) (Haley and Knott, 1992). For this analysis, WebQTL automatically does a permutation test to establish genome-wide significance criteria for the trait (Churchill and Doerge, 1994). By default, it returns a list of marker loci that show greater than suggestive association with the trait according to standard criteria (Lander and Kruglyak, 1995), but it will also accept user-defined criteria. Local maxima in the LRS in this list identify loci that are most likely to be near QTLs. WebQTL provides this list within a few seconds.

Since RI lines can provide multiple trait measurements per line, it is possible to estimate trait variance for each line separately. Variance of a trait can differ significantly among different RI lines. When variance estimates for each line are available, WebQTL uses these to improve the accuracy of the regression, and it provides an option for their submission.

Marker regression mapping evaluates potential QTLs only at the location of available marker loci. Once this mapping has identified

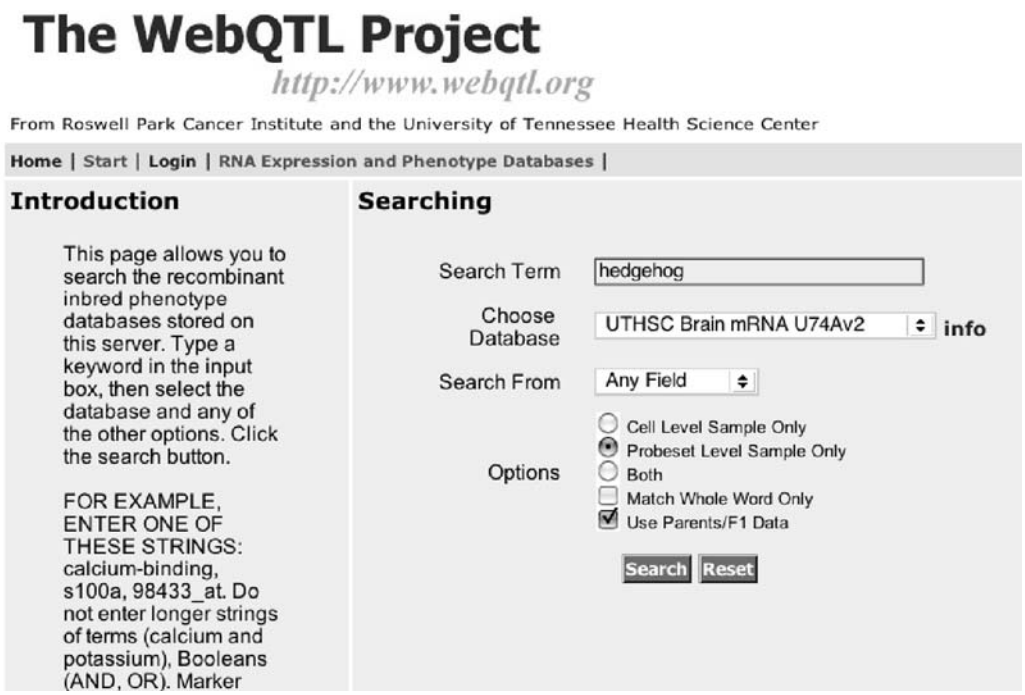


Fig. 2. Search dialog with request to retrieve data related to hedgehog from BXD brain gene expression data.

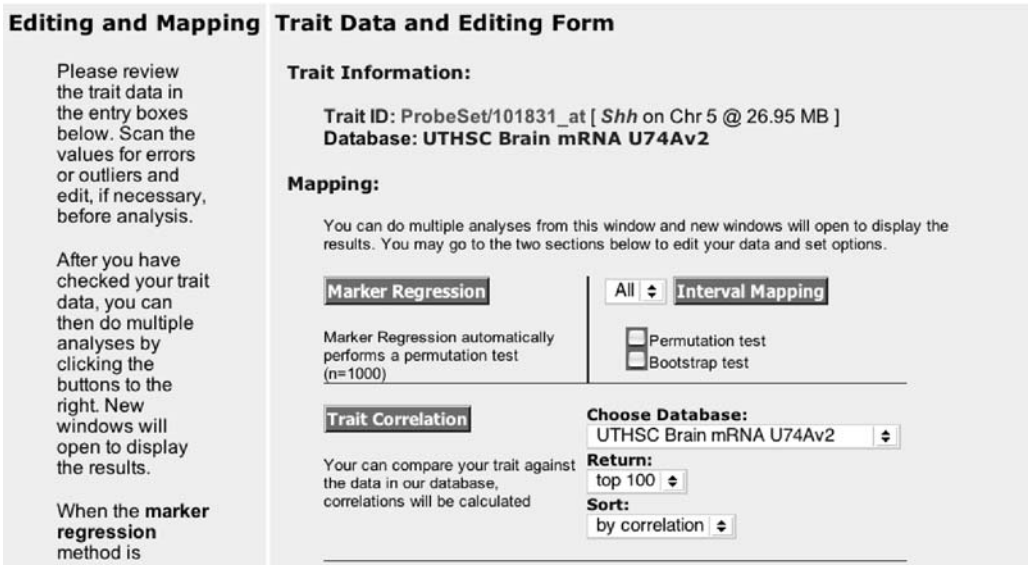


Fig. 3. Window with retrieved data ready for analysis. The data itself, not visible in the figure, is in an editable table reached by scrolling the window down. The buttons shown allow searches for correlations with other traits and allow QTL mapping by marker regression or interval mapping.

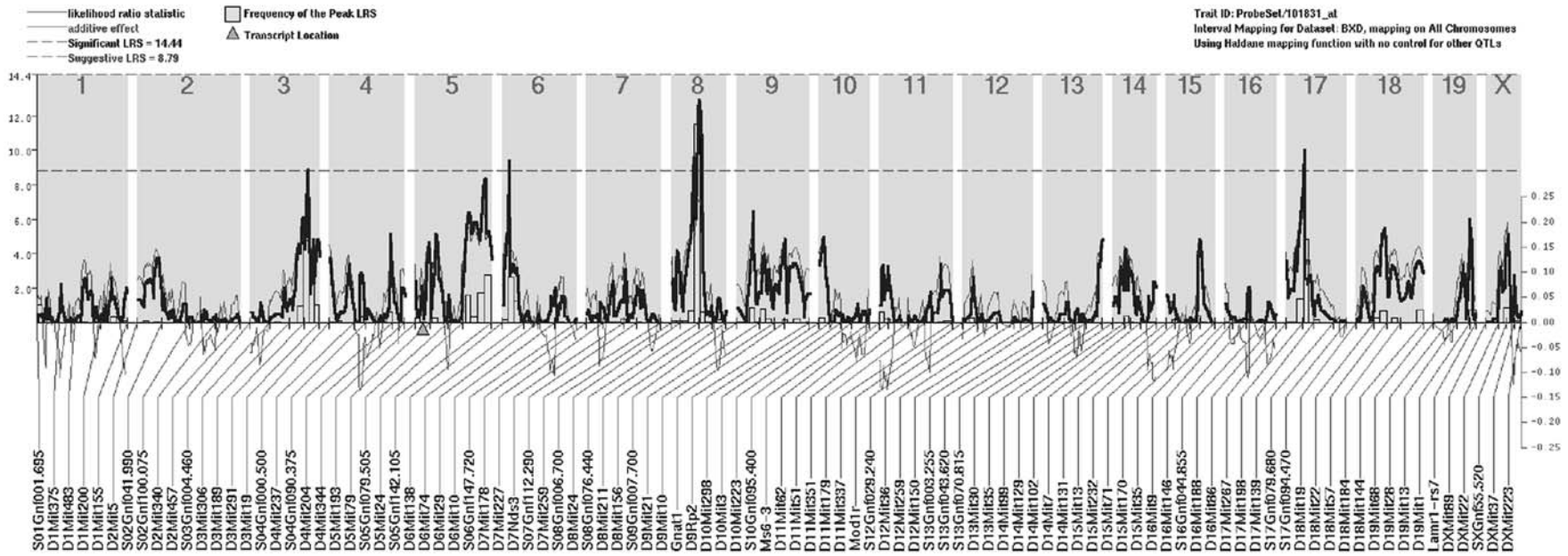


Fig. 4. Interval mapping of trait Shh brain RNA abundance on all mouse chromosomes. Peaks in the heavy line (LRS) show locations of a putative QTL, and the histogram beneath it shows frequent peak location for bootstrap samples. Permutation-based significance thresholds are given by dashed lines; the upper line corresponds to a genome-wide 5% significance threshold.

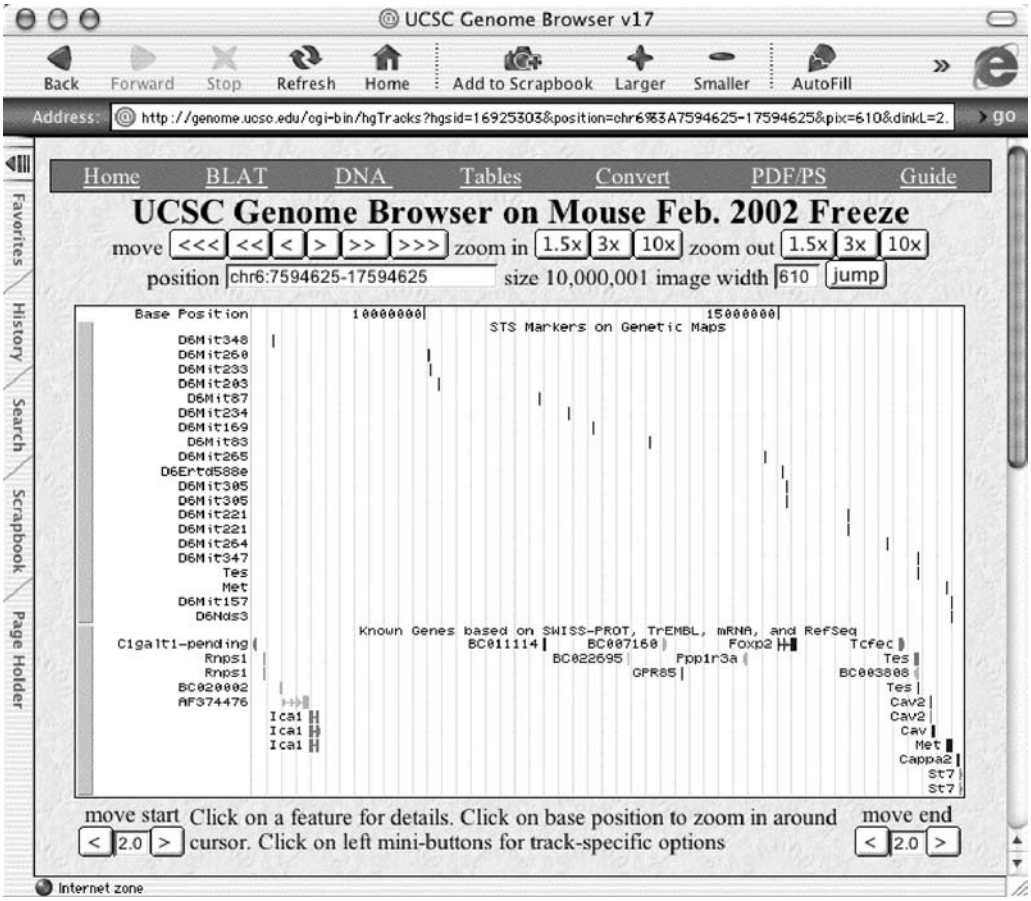


Fig. 5. The browser shows genomic features in the vicinity of the clicked position.

chromosomes with apparent QTLs, WebQTL provides interval mapping across single chromosomes (Haley and Knott, 1992). Interval mapping evaluates potential QTLs at regularly spaced intervals and estimates the significance at each location with an LRS, presenting the results graphically. Local maxima in the LRS curve identify the approximate location of QTLs. As a part of the interval mapping procedure, WebQTL offers an optional permutation test and an optional bootstrap analysis (Visscher et al., 1996). The permutation test establishes genome-wide criteria for significance (these may differ slightly from those for marker regression). The bootstrap test evaluates the reliability of the location suggested by an LRS maximum. It finds the location of the maximum LRS achieved across the chosen chromosome for 1000 bootstrap samples of the cross progeny, and displays a histogram showing the frequency at which the maximum appears at each chromosomal location. Figure 4 shows an example of interval mapping across the entire mouse genome, with permutation-based significance levels and bootstrap-based frequencies for peak location.

Positions in a WebQTL interval map are linked to the UCSC Genome Browser (Kent et al., 2002) to allow easy exploration of genes or predicted genes in the region of a QTL (Fig. 5).

For example, in Fig. 4 there is a potential QTL (not quite significant) at the proximal end of chromosome 6. In response to a mouse click on or under that peak, WebQTL will convert

the horizontal location of the click to a physical location on the chromosome and submit a request to http://genome.ucsc.edu. A new browser window will open, displaying a 10-mB region of the genome centered on the position of the mouse click.

When the first QTL has been identified for a trait, WebQTL offers the option of searching for a second while controlling for the effects of the first. Loci that achieved at least a suggestive LRS in a marker regression appear in a popup menu, and that with the highest LRS would generally be chosen to be a cofactor in the regression. This locus controls for the effect of any QTL(s) near it and may improve the power to detect a second locus. This type of search is done by marker regression; the corresponding option for interval mapping is not yet implemented.

QTL detection and mapping with microarray data involves, at two distinct levels, a statistical issue known as the "multiple testing problem." That is, QTL detection involves statistical tests of each trait at hundreds of positions in the genome, and QTL detection with microarray data may involve testing thousands of traits. The permutation tests described previously account for the tests at multiple genome positions. WebQTL does not attempt to correct for testing multiple traits because the current system does not allow automatic testing of multiple traits.

Data Resources

The WebQTL databases include genotype data for RI lines derived from about 1600 microsatellite loci recently collected, checked, and supplemented with new markers (Williams et al., 2001). Redundant markers were removed so that adjacent markers have at least one recombination between them. The number of markers is 499, 488, 756, 472, and 405, respectively, for the AXB, BXA, BXD, BXH, and CXB sets.

Trait data can also be retrieved from the integrated database, which contains data for two types of traits. First, it contains data collected from published literature for 274 biochemical and behavioral traits measured in BXD RI lines. Second, it contains over 800,000 traits defined by RNA abundance levels in BXD RI lines, measured by microarray hybridization with the Affymetrix U74Av2 GeneChip® (Lockhart et al., 1996). These data come from two sets of microarray measurements, one for brain (Chesler et al., 2003) and one for hematopoietic stem cells (Bystrykh et al., 2003). These data were specifically designed for discovery of QTLs controlling gene expression (Manly et al., 2002; Williams et al., 2002). Each set includes over 12,000 traits defined by average match-mismatch probe differences (Affymetrix Microarray Suite 5.0 algorithm), and almost 800,000 traits defined by individual match and mismatch probes. Trait data can be retrieved by searching the annotation associated with each trait. Annotation for the Affymetrix data includes the Affymetrix probe and probeset identifiers and may include a gene symbol and a short descriptive phrase (Figs. 1,2).

WebQTL includes genotypes for progeny of a tenth-generation BXD advanced intercross (Darvasi and Soller, 1995). These progeny were used to create a database of three-dimensional brain anatomy and histology that is available for study as stored images and as a web-accessible slide collection in the Mouse Brain Library (Rosen et al., 2003). This cross, therefore, is specifically designed for investigation of the genetics of brain anatomy and development. With these resources, quantitative trait loci can be defined by pure bioinformatics, by submitting to WebQTL traits defined by library images from the progeny of this cross.

Additional phenotype or RNA abundance data will be added to WebQTL as they become available.

Implementation

WebQTL is written largely in the Python and C languages. In particular, the CGI scripts of WebQTL are implemented using Python and several open-source Python modules called by those scripts (Numeric, PIL, PIDDLE, HTMLgen). To improve performance, a C language module for Python (qtl) was implemented to handle computation-intensive statistical analysis operations, such as regression, permutation, and bootstrap sampling. This module includes both matrix-based code for multiple regression and optimized code for simple regression and multiple regression with two independent variables. The integrated trait database is constructed using MySQL, an open-source database management system. A Python module (MySQLdb) allows Python functions to connect with and retrieve information from the MySQL database. Sources for the components mentioned are:

Python;
http://www.python.org/

Numeric;
http://sourceforge.net/projects/numpy

PIL;
http://www.pythonware.com

PIDDLE;
http://piddle.sourceforge.net/

HTMLgen;
http://starship.python.net/crew/friedrich/HTMLgen/html/main.html

MySQL;
http://www.mysql.com

MySQLdb;
http://sourceforge.net/projects/mysql-python

Acknowledgments

WebQTL is A Human Brain Project/Neuroinformatics program funded jointly by the National Institute of Mental Health, the National Institute on Drug Abuse, and the National Science Foundation (P20-MH 62009). We thank the following colleagues and collaborators for data resources included in WebQTL: Y. Qu, J. Gu, S. Qi, J. Hogenesch, R. Edwards, and L. Lu for genetic and genomic data; S. Shou, E.J. Chesler, L. Lu, H.C. Hsu, J.D. Mountz, and D.W. Threadgill for brain RNA expression data; G. de Haan, A. Su, and M. Cooke for hematopoietic stem cell RNA expression data; and E.J. Chesler and L. Lu for published biological and behavioral traits.

References

Barton, N. H. and Keightley, P. D. (2002) Understanding quantitative genetic variation. Nat. Rev. Genet. 3, 11–21.

Broman, K. W. (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab. Anim. (NY) 30, 44–52.

Bystrykh, L., Weersing, E., Sutton, S., et al. (2003) Genetical genomics to identify gene pathways in hematopoietic stem cells. Submitted.

Chesler, E. J., Wang, J., Lu, L., Qu, Y., Manly, K. F., and Williams, R. W. (2003) Genetic correlates of gene expression in recombinant inbred strains: A relational model system to explore neurobehavioral phenotypes. Neuroinformatics 1, 343–358.

Churchill, G. A. and Doerge, R. W. (1994) Empirical threshold values for quantitative trait mapping. Genetics 138, 963–971.

Darvasi, A. and Soller, M. (1995) Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141, 1199–1207.

Doerge, R. W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. 3, 43–52.

Haley, C. S. and Knott, S. A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69, 315–324.

Haley, C. S., Knott, S. A., and Elsen, J.-M. (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136, 1195–1207.

Kent, W. J., Sugnet, C. W., Furey, T. S., et al. (2002) The human genome browser at UCSC. Genome Res. 12, 996–1006.

Knott, S. A., Elsen, J.-M., and Haley, C. S. (1996) Methods for multiple marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. 93, 7180.

Lai, K., Kaspar, B. K., Gage, F. H., and Schaffer, D. V. (2003) Sonic hedgehog regulates adult neural progenitor proliferation in vitro and in vivo. Nat. Neurosci. 6, 21–27.

Lander, E. and Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. 11, 241–247.

Lockhart, D. J., Dong, H., Byrne, M. C., et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol. 14, 1675–1680.

Mackay, T. F. C. (2001) The genetic architecture of quantitative traits. Annu. Rev. Genet. 35, 303–339.

Manly, K. F., Wang, J., Shou, S., et al. (2002) QTL mapping with microarray expression data. In: 16th International Mouse Genome Conference, San Antonio, TX.

Mott, R. and Flint, J. (2002) Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. Genetics 160, 1609–1618.

Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C., and Flint, J. (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA 97, 12649–12654.

Phillips, T. J. and Belknap, J. K. (2002) Complex-trait genetics: emergence of multivariate strategies. Nat. Rev. Neurosci. 3, 478–485.

Rosen, G. D., La Porte, N. T., Diechtiareff, B., et al. (2003) Informatics center for mouse genomics: The dissection of complex traits of the nervous system. Neuroinformatics 1, 327–342.

Seaton, G., Haley, C. S., Knott, S. A., Kearsey, M., and Visscher, P. M. (2002) QTL Express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics 18, 339–340.

Seaton, G. G., Kearsey, M. J., Snape, J. W., Haley, C. S., and Nudd, G. R. (1998) JAVA, A programming solution to multi-platform analysis using regression techniques. In: Plant & Animal Genome VI, San Diego, CA.

Threadgill, D. W., Airey, D. C., Lu, L., Manly, K. F., and Williams, R. W. (2001) Recombinant inbred intercross (RIX) mapping: A new approach extending the power of existing mouse resources. In: 15th International Mouse Genome Conference, p. 50, Edinburgh, UK.

Visscher, P. M., Thompson, R., and Haley, C. S. (1996) Confidence intervals in QTL mapping by bootstrapping. Genetics 143, 1013–1020.

Williams, R. W., Gu, J., Qi, S., and Lu, L. (2001) The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. Genome Biol. 2, research0046.0041-0046.0018.

Williams, R. W., Manly, K. F., Shou, S., et al. (2002) Massively parallel complex trait analysis of transcriptional activity in mouse brain. In 16th International Mouse Genome Conference, San Antonio, TX.